VFMM3D：単眼3D物体に対する視覚基盤モデルによる画像の可能性の解放

Detection

Bonan Ding, Jin Xie, Jing Nie, Jiale Cao, Xuelong Li, and Yanwei Pang

概要-単眼3D物体検出は、その費用対効果と広範な利用可能性により、推論時に単一のカメラのみに依存するため、自律走行やロボット工学を含む様々なアプリケーションにおいて重要な意味を持つ。とはいえ、単眼画像から3次元空間における物体の座標を直接予測することは困難である。したがって、効果的なソリューションは、単眼画像をLiDARのような表現に変換し、オブジェクトの3D座標を予測するためにLiDARベースの3Dオブジェクト検出器を採用することである。この方法の重要なステップは、単眼画像を信頼性の高い点群形式に正確に変換することである。本論文では、Vision Foundation Models (VFM)の機能を活用し、シングルビュー画像をLiDAR点群表現に正確に変換する革新的なフレームワークであるVFMM3Dを紹介する。VFMM3Dは、Segment Anything Model (SAM)とDepth Anything Model (DAM)を利用し、豊富な前景情報でリッチな高品質の擬似LiDARデータを生成する。具体的には、Depth Anything Model (DAM)を用いて、高密度の深度マップを生成する。その後、セグメント何でもモデル(SAM)を利用して、インスタンスマスクを予測することで、前景と背景の領域を区別する。これらの予測されたインスタンスマスクと深度マップを組み合わせて3D空間に投影し、擬似LiDAR点を生成する。最後に、点群に基づくあらゆる物体検出器は、物体の3次元座標を予測するために利用することができる。KITTIとWaymoの2つの困難な3D物体検出データセットを用いて包括的な実験を行った。我々のVFMM3Dは、両データセットにおいて、新たな最先端性能を確立した。さらに、実験結果はVFMM3Dの汎用性を実証し、様々なLiDARベースの3Dオブジェクト検出器へのシームレスな統合を示す。

索引用語-3D物体検出、視覚基盤モデル、単眼視覚。

I. Introduction

自律システムの台頭は、環境を3つの次元で認識・理解することが単に有利であるだけでなく、必須である時代をもたらした。様々なセンシング手法の中で、単眼視はその単純さと費用対効果の高さから、魅力的な選択肢として浮上している。しかし、1枚の2D画像から3D情報を抽出することは困難であり、大きな障害となる。単眼3D物体検出[1]-[5]は、単眼画像から物体の3Dバウンディングボックスを推定することを任務とし、特に高度なドライバ支援システム(ADAS)、ロボット工学、バーチャルリアリティにまたがる数多くのアプリケーションにおいて極めて重要な要素である。

従来の単眼3D物体検出アプローチは、ジオメトリベースの手法[6]-[11]や奥行き推定技術[12]-[18]に依存しており、特に複雑で動的なシーンでは、精度とロバスト性の欠如に悩まされることが多い。

最近のディープラーニングの急増は、畳み込みニューラルネットワーク(CNN)と視覚変換器(ViT)[19]によって新しい可能性をもたらし、コンピュータビジョン分野に革命をもたらした。これらのモデルは、広範なデータセットで事前に訓練され、多数のタスクにわたって汎化し、良好な性能を発揮する不気味な能力を実証している。

Depth Anything Model (DAM) [20]やSegment Anything Model (SAM) [21]などの視覚基盤モデル(VFM)は、単眼3D物体検出の領域で特に有望視されている。正確なセグメンテーションマスクを生成する能力を持つSAMと、単眼画像から正確なシーン深度を推定する能力を持つDAMは、単眼3D検出モデルの性能を大幅に向上させることができる深度特徴の豊富なセットを提供する。これらのVFMの組み合わせは、奥行きの曖昧さやロバストなセグメンテーションの必要性など、単眼3D検出に内在する課題に対処するユニークな機会を提供する。

本論文では、SAMとDAMを相乗的に統合し、単眼3D物体検出を精密に行う新しいフレームワークであるVFMM3Dを紹介する。我々のアプローチは、意味情報と正確な深度で強化された擬似LiDARデータを生成するために、両モデルの長所を活用する。本フレームワークの能力は、将来、より強力なビジョン基盤モデルやLiDARベースの3Dオブジェクト検出器によって、さらに向上する可能性がある。生成された高品質の擬似LiDARデータに基づき、最先端のLiDARベースの3D物体検出器を採用して3D物体検出を行うことができる。

VFMM3Dは、単眼3D検出フレームワークにいくつかの革新的なコンポーネントを導入している。まず、特定のデータセットで微調整されることに依存せず、高品質で前景情報に富んだ擬似LiDARデータを生成する方法を提案し、広く適用できるようにする。次に、密な擬似LiDAR点に関連する計算効率の悪さとノイズに対処するために、スパース化技術を組み込む。最後に、本手法は様々なLiDARベースの3D物体検出器に適応し、その汎用性と柔軟性を示している。

KITTIデータセットでの広範な実験を通して、我々はVFMM3Dが異なる難易度において、既存の最先端の単眼3D物体検出手法を凌駕することを実証する。VFMM3Dの優れた性能は、VFMによって単眼画像から詳細な3次元空間情報を効果的に抽出・利用できることに起因する。

我々の研究は、単眼3D物体検出のフロンティアを押し広げるだけでなく、実世界のアプリケーションに容易に適用できる、ロバストで一般化可能なソリューションを提供する。結論として、貢献度は以下の通りである:

- 我々の知る限り、VFMM3Dは視覚基礎モデルと単眼3D物体検出タスクを統合した最初のアプローチである。VFMM3Dは、特定のデータセット上で微調整を行うことなく、SAMとDAMを利用し、任意のシーンにおける単眼3D物体検出のためのLiDARベースの3D検出器と組み合わせる。
- 本手法で導入した擬似LiDAR絵画操作により、2D画像が3Dタスクに提供できる3D情報を十分に活用し、3D空間におけるSAMとDAMの結果をより良く統合することができ、最終的な検出精度を大幅に向上させることができる。
- VFMM3Dは、視覚基礎モデルの仮想点生成と任意の3D物体検出器とのシームレスな統合を可能にするスパース化操作を導入している。検出精度を大幅に向上させ、計算コストと推論時間を大幅に削減する。
- KITTIとWaymoデータセットを用いた広範な実験により、図2に示すように、本手法は既存の単眼3d物体検出手法の中で最先端の結果を達成していることが示された。

II. Related Work

視覚基盤モデル。視覚基盤モデル(VFM)は、コンピュータビジョンの領域における重要な進歩であり、広範なデータセットに対する頑健な事前学習により、様々なタスクにわたって汎用性の高いソリューションを提供する。VFMの中でも、Vision Transformers (ViTs) [19]は、LVD142M [22]のような巨大なデータセットで学習された、極めて重要なモデルである。ViTの有効性は、DINO [23]やDINOv2 [22]のような、知識蒸留技術と組み合わせた自己教師あり学習を活用するアプローチによって、さらに増幅される。VFMの顕著な応用の1つは、Segment Anything Model (SAM) [21]に見られる。SAMは、画像内の個々の要素に対して正確なマスクを生成することに長けており、詳細なオブジェクトのセグメンテーションを扱う能力を示している。1100万枚の画像と11億枚のマスクを含む膨大なデータセットSA-1Bで学習している。同様に、Depth Anything Model (DAM) [20]は、単眼奥行き推定のためのロバストなソリューションとして登場し、2D画像を3D空間に正確に投影することを可能にする。この機能は、画像内の各ピクセルに奥行き情報を提供するのに役立ち、シーンジオメトリのより包括的な理解を促進する。VFMM3Dモデル内のSAMとDAMの統合は、単眼3D物体検出を強化するための新しいアプローチである。SAMのセグメンテーション能力は、擬似LiDARデータ内の前景情報を洗練するために利用され、DAMの深度推定能力は、2D画像を3D空間に投影するのに役立ち、それによって2D画像から可能な限り多くの3D空間表現のための貴重な情報を抽出する。

単眼3D物体検出。単眼3D物体検出の分野では、1枚の2D画像から3D情報を抽出する様々な手法が開発されている。Deep3DBox[8]は、ヨー推定のためのMultiBinアプローチを導入し、2Dバウンディングボックスの幾何学的制約を利用して、3Dバウンディングボックスを生成する。MonoPair[6]は、オブジェクト間の空間的関係を利用して3D特性を推定する、典型的なジオメトリベースの手法である。MonoFlex[24]は、切り捨てられたオブジェクトを分離する柔軟なフレームワークを提案し、直接回帰やキーポイントからの幾何学的解を含む複数の深度推定アプローチを統合している。MonoDLE[16]とPGD[25]は、ディープラーニングを適用して、画像からインスタンスの3Dバウンディングボックスや奥行きを直接予測する。MonoRUn[26]は、3Dバウンディングボックスの注釈を直接使用することで、密な対応関係と幾何学的関係を確立するために、自己教師あり学習を活用する。MonoDTR[27]とMonoDETR[28]は、単眼3D物体検出のための深度ガイド変換器を介した適応的な特徴集約を行う。擬似LiDARベースの手法[29]-[32]は、画像から擬似LiDAR点群を生成し、ボクセルベースの技術を用いて処理する。CaDDN [17]は、各ピクセルのカテゴリカルな深度分布を学習することで、鳥瞰図(BEV)表現を構築する。そして、BEV投影からバウンディングボックスを復元する。MonoPSR [7]は、インスタンス点群を推定し、オブジェクトの外観と投影点群との間の整合を強制することで、提案を洗練する。QNet[33]は擬似LiDARデータを画像表現に変換し、強力な2次元CNNを用いて検出性能を向上させる。この問題に取り組むために、M3D-RPN [12]は、抽出された特徴を強化するために、3D領域提案ネットワーク内に深度を考慮した畳み込み層を導入する。RTM3D [34]は、3Dバウンディングボックスの投影頂点を推定し、非線形最小二乗最適化により3D特性を解決する。本論文では、擬似LiDARベースのフレームワークを採用し、その優れた一般性を考慮し、多様なシナリオに対応するために、様々な検出器とのシームレスな統合を可能にする。LiDARベースの3D物体検出。LiDARベースの3Dオブジェクト検出の領域では、ポイントベースとボクセルベースの2つの主要なパラダイムがある。PointRCNN[35]は前者を例示し、PointNet++[36]を利用する、

点ベース手法のパイオニアであり、生のLiDAR点群からの特徴抽出に続き、[37]-[39]が続く。これとは対照的に、VoxelNet [40]は、3D LiDAR点群を等間隔に配置された3Dボクセルに変換する、ロバストなボクセル特徴エンコーディング層を示す。このボクセル中心のアプローチに基づき、SECOND [41]はLiDAR点群用に調整された新しいスパース3D畳み込み層を導入している。従来のボクセル手法とは異なり、PointPillars [42]は特徴量の符号化に柱を採用し、ネットワーク効率を向上させている。同様に、VoxelRCNN [43]は、ボクセルの特徴を2段階のフレームワークで完全に利用するために、ボクセルクエリとボクセルROIプーリング法を提案している。PV-RCNN[44]は、点とボクセルの利点を統合し、ボクセルセットの抽象化を採用して、マルチスケールボクセル特徴をサンプリングされたキーポイントに融合させる。

III. Method

A. Task Definition

単眼3D物体検出の目的は、単一のビューRGB画像とそれに対応するカメラパラメータを与えることによって、

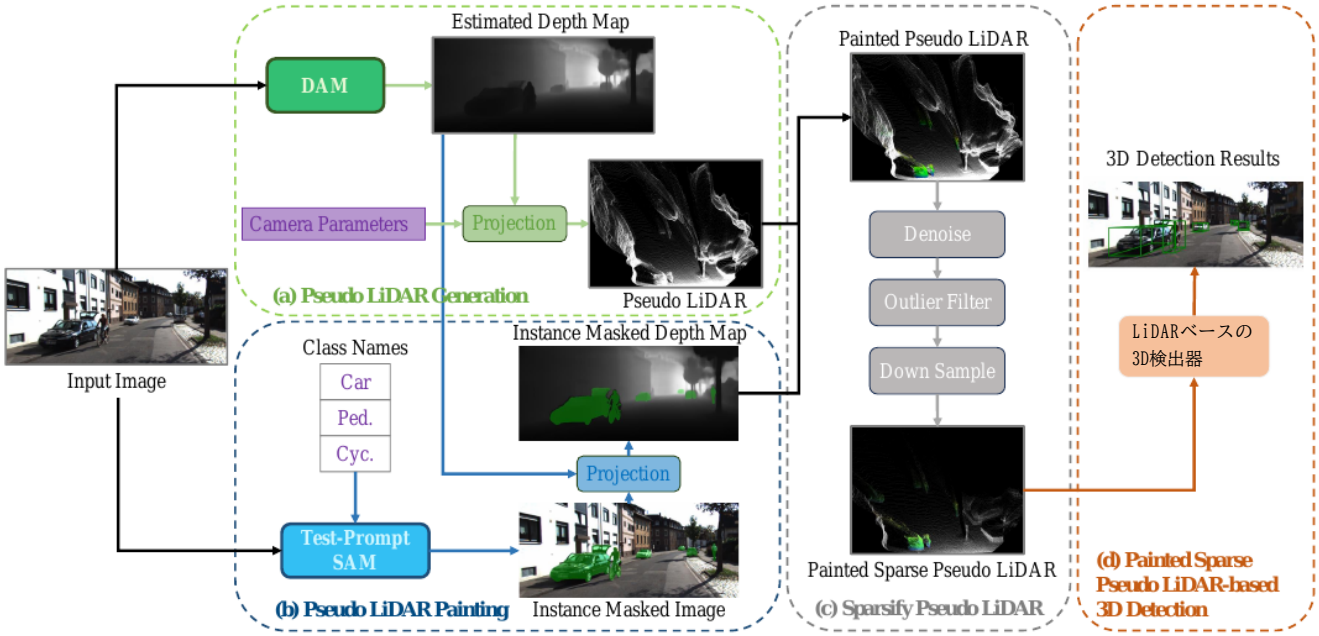

図1. VFMM3Dの全体アーキテクチャ。我々のモデルは、(a)DAMによる擬似LiDAR生成、(b)テキストプロンプトSAMによる擬似LiDAR絵画、(c)擬似LiDARスパース化、(d)LiDARベースの検出器による3D物体検出の4つの部分から構成される。

3D空間における関心物体を分類し、位置を特定することである。各オブジェクトは、その中心座標(x, y, z)、サイズ(h, w, l)、および方向θによってパラメータ化されたカテゴリスコアとバウンディングボックス(BBox)によって表現される。

B. フレームワークの概要

擬似LiDARに基づく既存の単眼3D物体検出手法は、多くの場合、事前に訓練された奥行き推定と、特定のデータセットで事前に訓練されたセマンティックセグメンテーションモデルを必要とする。特定のデータセットでの事前学習を必要とせず、あらゆるデータセットに適用可能な高品質の擬似LiDARデータを生成するために、DAMとSAMという2つの視覚基盤モデルを活用したVFMM3Dを紹介する。DAMは、任意のデータセットとシーンに対して、ロバストな擬似LiDARを生成するタスクを課される。しかし、実際のLiDARデータと比較すると、この深度予測によって生成された擬似LiDARはまだ精度に欠ける。この問題に対処するため、SAMを統合して擬似LiDARに正確な前景情報を提供し、不正確な深度推定に起因する不正確さを緩和する。さらに、従来のLiDARベースの3Dオブジェクト検出器は、疎な点群データ用に設計されている。検出速度を向上させ、冗長で不正確な擬似LiDAR点による潜在的な精度への影響を緩和しながら、既存の検出器に適応するために、効率と精度を向上させることを目的としたスパース化手法を提案する。VFMM3Dの全パイプラインを図1とアルゴリズム1に示すが、主に4つのステップを含む:

- 擬似LiDAR生成。このステップでは、まずDAMを用いて深度マップを取得する。次に、この深度マップを3次元空間に投影し、擬似LiDARデータを得る。
- **Pseudo-LiDAR Painting**.

Algorithm 1 VFMM3D

Input:
RGB image $I \in \mathbb{R}^{H\times W\times 3}$.

Output:
Predicted 3D bounding boxes $B_{3D} = \{(x_i, y_i, z_i, h_i, w_i, l_i)\}_{i=1}^{M}$

$D = Depth\ Generation(I)$ ▷ Depth Map $D \in \mathbb{R}^{H\times W\times 1}$
$P = Projection(D)$ ▷ Pseudo-LiDAR $P \in \mathbb{R}^{N\times 3}$
$I_M = Mask\ Generation(I, B)$ ▷ Instance Mask $I_M \in bool^{H\times W}$
$\hat{P} = Painting(D, I_M, P)$ ▷ Painted Pseudo-LiDAR $\hat{P} \in \mathbb{R}^{N\times 6}$
$\bar{P} = Sparsification(P)$ ▷ Sparse Painted Pseudo-LiDAR $\bar{P} \in \mathbb{R}^{\bar{N}\times 6}$
$B_{3D} = LiDAR\text{-}Based\ 3D\ Detector(\bar{P})$

SAMはRGB画像を入力とし、テキストプロンプトを組み合わせて前景オブジェクトのセグメンテーションを実行する。前景オブジェクトのセグメンテーション結果は、DAMによって得られた深度マップにマッピングされ、前景オブジェクトの深度マップを強調し、ノイズをフィルタリングするために採用される。そして、対応する擬似LiDARもより正確である。

- 擬似LiDARスパース化。深度マップから得られる擬似LiDAR点の数は、実際のシーンにおけるLiDAR点群よりもはるかに多いため、最終的なLiDARベースの3dオブジェクト検出器に適応するために、ペイントされた擬似LiDARを適切にスパースするスパース化ステップを採用する。

- LiDARベースの3D検出。ペイントされた擬似LiDARを用いて3D物体検出を行うために、LiDARベースの3D物体検出器を採用する。この検出器は、前のステップで生成されたペイントされた擬似LiDARを入力とし、最終的な検出結果を生成する。

C. DAMによる擬似LiDAR生成

前置き。我々の単眼3D物体検出パイプラインの基礎となるステップは、単眼画像からの奥行き推定である。ロバストな単眼奥行き推定(MDE)のために、Depth Anything Model(DAM)[20]を導入する。DAMは技術的なモジュールに頼らず、様々な公開大規模データセットから約6200万枚のラベルなし画像を自動的に収集し、注釈を付けるように設計された新しいデータエンジンでデータセットをスケールアップすることに焦点を当てている。ラベル付けされていないデータを広範囲に使用することで、データカバレッジが大幅に広がり、それによって汎化誤差が減少し、多様で困難なシーンを扱うモデルの能力が向上する。

DAMは、そのパフォーマンスを向上させるために、2つの戦略を活用している。第一に、データ補強ツールの使用により、より困難な最適化ターゲットを作成し、モデルが能動的に追加的な視覚的知識を求め、ロバストな表現を獲得するよう促す。第二に、事前に訓練されたエンコーダから豊富な意味的事前分布を継承するようにモデルを強制する補助的な監視メカニズムを組み込んでいる。上記のアプローチは、奥行き推定におけるモデルの性能を向上させるだけでなく、中位と上位の両方の知覚タスクを処理できるマルチタスクエンコーダを提供する。

DAMによる深度推定。DAMはさらに、NYUv2やKITTIなどの標準的なデータセットからメトリック深度情報を用いて微調整され、メトリック深度推定精度において新たな最先端記録が設定される。単一の画像$I \in R^{H\times W\times 3}$が与えられると、DAMによって生成された信頼できる深度マップ$D \in R^{H\times W\times 1}$を得ることができる。深度マップを得た後、それを3次元空間に投影し、実世界座標$P = \{(x^{(n)}, y^{(n)}, z^{(n)})\}_{n=1}^{N}$の擬似LiDARを、深度マップをカメラ座標系に最初に投影して求める：

$$\begin{cases} z_c = d, \\ x_c = \frac{(u-C_x)\times z}{f}, \\ y_c = \frac{(v-C_y)\times z}{f}, \end{cases} \tag{1}$$

ここで、dは深度マップにおける画素(u, v)の推定深度、(C_x, C_y)はカメラの主点である。f_xとf_yはそれぞれx軸とy軸に沿ったカメラ焦点距離である。Nは画素数である。そして、カメラ外部行列M_Eが与えられたとき：

$$M_E = \begin{bmatrix} R & t \\ 0^T & 1 \end{bmatrix}, \tag{2}$$

ここでM_Eは4×4行列である。Rは3×3直交単位行列で、回転行列としても知られている。t は3次元の並進ベクトルである。そして、ワールド座標系における擬似LiDARは次式で得られる：

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = M_E^{-1} \begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} \tag{3}$$

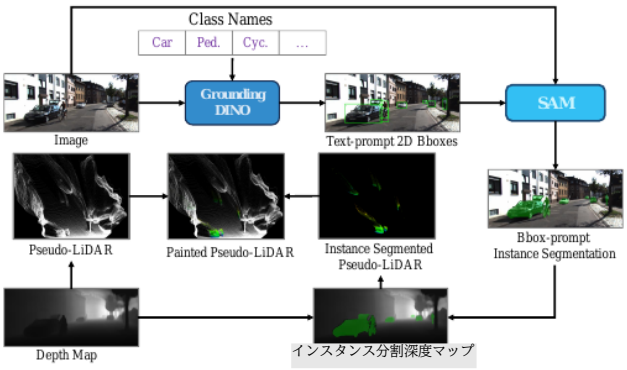

図2. 擬似LiDARペインティングテキストプロンプトSAMの詳細アーキテクチャ。

D. テキストプロンプトSAMによる擬似LiDARペインティング

予備知識 Segment Anything Model (SAM) [21]は、1100万枚以上の画像と10億枚の驚くべきマスクからなる広範なSA-1Bデータセットで学習されたVFMとしてのコンピュータビジョンの分野における大きな飛躍であり、異なるシーンやオブジェクトタイプにまたがる顕著な汎化能力を付与している。SAMの設計は、軽量マスクデコーダと強力な画像エンコーダに支えられており、これらは同時にプロンプトを処理し、高品質なセグメンテーション出力を生成する。そのプロンプト可能な性質は、点、バウンディングボックス、さらには自由形式のテキストを含む様々な形式の入力を柔軟に受け入れることを可能にし、多くのアプリケーションでアクセス可能にする。SAMファミリーの手法[45]、[46]は、この基礎をさらに発展させ、SAMのコア機能を強化する様々な構成と適応を探求している。これらの手法は広範囲に評価され、印象的なゼロショット性能と、様々なセグメンテーションタスクにおいて完全教師ありモデルと競合する、あるいはそれを上回る能力を示している。SAMとそのファミリー手法は、2Dと3Dの空間表現の間に固有のギャップがあるため、3Dシーンに直接適用することはできない。しかし、これらを単眼3D物体検出に適用することで、このタスクのためのロバストで強力な意味レベルの画像特徴を自然に提供することができる。

テキストプロンプトSAMによる深度マップのセグメンテーション。PointPainting[47]、MonoPseudo[31]、Fusionpainting[48]で述べられているように、セマンティックセグメンテーションによって画像中の前景オブジェクトから抽出されたセマンティック情報は、3Dオブジェクト検出の精度とロバスト性を高める上で重要な役割を果たす。本手法の幅広い適応性を示すために、自律走行データセットと追加的な事前データで特別に事前学習された2次元物体検出モデルの使用を回避する。その結果、点、バウンディングボックス、マスクなど、普段のような手がかりをSAMに提供することができない。代わりに、このアプローチで使用されるSAMは、テキストベースのプロンプトに基づいている。具体的には、テキストプロンプトSAMは、図2に示すように、テキストプロンプトオープンセット2dオブジェクト検出器Grounding DINO [49]とSAMの2つの部分から構成される。

Algorithm 2 Pseudo-LiDAR Painting with Text-Prompt SAM

Input:
RGB image $I \in \mathbb{R}^{H\times W\times 3}$.
Depth map $D \in \mathbb{R}^{H\times W\times 1}$.
Pseudo-LiDAR points $P \in \mathbb{R}^{N\times 3}$ projected from D.

Output:
Painted pseudo-LiDAR points $\hat{P} \in \mathbb{R}^{N\times 3+3}$

$B_{2D} = GroundingDINO(I)$ ▷ $B_{2D} \in \mathbb{R}^{K\times 4}$
$I_M = SAM(I, B_{2D})$ ▷ $I_M \in bool^{H\times W}$
$D_M = D \times I_M$

for $d \in D_M, p \in P$ **do**
  if $d \neq 0$ **then**
    $i = I[u, v, :]$ ▷ $i \in \mathbb{R}^3$
    $\hat{p}$ =Concatenates(p, i) ▷ $\hat{p} \in \mathbb{R}^6$
  else
    $\hat{p}$ =Concatenates$(p, \mathbf{0}_3)$ ▷ $\hat{p} \in \mathbb{R}^6$
  end if
end for

まず、グラウンディングDINOはRGB画像Iを取り込み、2次元物体検出BBox B_{2D} = $\{(u_B^{(k)}, v_B^{(k)}, h_B^{(k)}, l_B^{(k)})\}_{k=1}^{K}$ をテキストプロンプト(すなわち、. ここで、Kは予測BBoxの数、u_B，v_B，h_B，l_Bは中心画素座標を表し、thはe 各予測BBoxのサイズをそれぞれ示す。次に、RGB画像Iと予測された2d BBoxes B_{2D}をSAMに入れ、BBoxes B_{2D}をプロンプトとして設定し、セマンティックセグメンテーションマスク画像I_Mを得る。次に、I_MのマスクをDAMから生成された対応する深度マップDに投影し、前景マスクされた深度マップD_Mを得る。Pointpainting[47]に触発されて、D_MのピクセルのRGBチャンネルを対応する擬似LiDAR Pに追加し、ペイントされた擬似LiDAR P^を得る。ペインティングアルゴリズムの詳細をアルゴリズム2に示す。

E. 擬似LiDARのスパース化

最近の多くの単眼[29]、[31]、[32]やマルチモーダル融合ベース[50]-[52]の3D物体検出器は、奥行き推定や補完アルゴリズムによって画像から生成された擬似LiDARを利用している。しかし、擬似LiDARは、画像中の画素数と同数の擬似点を持つため、スキャンで捉えた真のLiDARに比べてはるかに高密度である。さらに、VirConv [52]で述べられているように、オリジナルの擬似LiDARは、検出速度の低下(真のLiDARベースの3Dオブジェクト検出器よりも2 3倍以上遅い)につながる、かなりの計算コストを導入している。さらに、LiDARベースの3Dオブジェクト検出器では、前景オブジェクトに対応する点のみがフォーカスされるため、背景点が多すぎると、前景オブジェクトからネットワークが散漫になる可能性がある。上記の問題以外にも、擬似LiDARはノイズの問題にも悩まされる。これらのノイズは深度推定の不正確さに起因し、非ガウス分布ノイズとなる。従来の方法では、このようなノイズを除去するのに苦労していた。Pseudo-LiDAR++[50]では、安価な2ビームの真のLiDAR点群を擬似LiDAR補正に使用し、これらのノイズを安価に軽減する。しかし、我々の提案する方法は単一画像データのみに依存する。そこで、これらのノイズの影響を可能な限り最小化するために、いくつかの伝統的な方法を採用する。スパース化操作を図1に示す。具体的には、ペイントされた擬似LiDAR P^を得た後、まずP^の各点を球座標空間にマッピングし、この球座標空間の座標を用いてP^を小さなボクセルサイズでボクセル化する。各ボクセル内の全点の平均3次元座標のみを保持し、ノイズ除去されたペイント擬似LiDAR $P^{\wedge}_{d}$ を得る。次に、$P^{\wedge}_{d}$の外れ値擬似点を、あらかじめ定義された範囲に基づいてフィルタリングする。具体的なフィルタリング範囲は、データセットによって異なる設定となるが、これについてはセクションIVで詳述する。最後に、より大きなボクセルサイズを用いて擬似LiDARの3次元座標に基づいてボクセル化し、各ボクセル内の点が固定点数以下になるようにランダムにサンプリングする(この方法では5点)ことで、最終的にスパースペイント擬似LiDAR $P^{-} \in^{N\times 6}_{R}$, $N^{-} \ll N$ を得ることができる。

F. LiDARベースの検出

ペイントされた擬似LiDAR P^{-}は、任意のLiDARベースの3Dオブジェクト検出器に入力することができ、3Dオブジェクトの結果を得ることができる。本研究では、LiDARのみの3D検出器に焦点を当てるが、本手法では、マルチモーダル融合(すなわち、LiDAR+画像の入力)ベースの3D検出器を使用することができる。この決定は、検出速度のために、擬似LiDAR点数が実際の点数に比べて著しく高いことを考慮して行われる。マルチモーダル融合ベースの検出器を採用することで、検出速度が大幅に低下する。したがって、本研究では単一のLiDARベースの3D検出器のみを考慮する。

具体的には、VFMM3Dが3つの異なるLiDARベースの検出器と互換性があることを示す: PV-RCNN[44]、VoxelRCNN[43]、PointPillars[42]である。これらは広く使われているLiDAR検出器を含み、それぞれがシングルステージ(PointPillars)とツーステージ(PV-RCNNとVoxel-RCNN)という異なるネットワークアーキテクチャを特徴としている。

Voxel-RCNN. The Voxel-RCNN method is designed to enhance the performance of voxel-based 3D object detection while maintaining computational efficiency. It introduces a two-stage framework that includes a 3D backbone network for feature extraction, a 2D bird-eye-view (BEV) Region Proposal Network (RPN), and a detection head. The key innovation is voxel RoI pooling, which straightforwardly extracts RoI features from voxel features for subsequent refinement. This approach allows for real-time frame processing while achieving comparable detection accuracy to that of state-of-the-art point-based approaches, significantly reducing computation costs.

PV-RCNN PV-RCNN (Point-Voxel Feature Set Abstraction for 3D Object Detection)は、3Dボクセル畳み込みニューラルネットワークとPointNetベースの集合抽象化を深く統合した高性能な3D物体検出フレームワークである。3Dボクセル CNNの効率的な学習により高品質な提案を実現し、PointNetベースのネットワークの柔軟な受容野を利用して、正確な位置情報とコンテキストを捉えることができる。

TABLE I

COMPARSION RESULT ON WAYMO *val.* SET FOR THE VEHICLES CLASS. ボクセルではなくピラーを使用することで、PointPillarsはハンドR EDが最良の結果を示す必要性を排除し、一方、Blueは2番目に良い結果を示す。

Diff.	Extra	Method	Reference	3D mAP / mAPH(IoU=0.7)				3D mAP / mAPH(IoU=0.5)			
				Overall	0 - 30m	30 - 50m	50m - ∞	Overall	0 - 30m	30 - 50m	50m - ∞
L1	LiDAR	CaDDN [17]	CVPR 21	5.03 / 4.99	14.54 / 14.43	1.47 / 1.45	0.10 / 0.10	17.54 / 17.31	45.00 / 44.46	9.24 / 9.11	0.64 / 0.62
	LiDAR	MonoNeRD [53]	ICCV 23	10.66 / 10.56	27.84 / 27.57	5.40 / 5.36	0.72 / 0.71	31.18 / 30.70	61.11 / 60.28	26.08 / 25.71	6.60 / 6.47
	LiDAR	DID-M3D [54]	ECCV 22	- / -	- / -	- / -	- / -	20.66 / 20.47	40.92 / 40.60	15.63 / 15.48	5.35 / 5.24
	Depth	PatchNet [55]	ECCV 20	0.39 / 0.37	1.67 / 1.63	0.13 / 0.12	0.03 / 0.03	2.92 / 2.74	10.03 / 9.75	1.09 / 0.96	0.23 / 0.18
	Depth	PCT [56]	NeurIPS 21	0.89 / 0.88	3.18 / 3.15	0.27 / 0.27	0.07 / 0.07	4.20 / 4.15	14.70 / 14.54	1.78 / 1.75	0.39 / 0.39
	-	M3D-RPN [12]	ICCV 19	0.35 / 0.34	1.12 / 1.10	0.18 / 0.18	0.02 / 0.02	3.79 / 3.63	11.14 / 10.70	2.16 / 2.09	0.26 / 0.21
	-	GUPNet [9]	ICCV 21	2.28 / 2.27	6.15 / 6.11	0.81 / 0.80	0.03 / 0.03	10.02 / 9.94	24.78 / 24.59	4.84 / 4.78	0.22 / 0.22
	-	DEVIANT [57]	ECCV 22	2.69 / 2.67	6.95 / 6.90	0.99 / 0.98	0.02 / 0.02	10.98 / 10.89	26.85 / 26.64	5.13 / 5.08	0.18 / 0.18
	-	MonoJSG [58]	CVPR 22	0.97 / 0.95	4.65 / 4.59	0.55 / 0.53	0.10 / 0.09	5.65 / 5.47	20.86 / 20.26	3.91 / 3.79	0.97 / 0.92
	-	SSD-MonoDETR [59]	TIV 2023	4.54 / -	9.93 / -	1.18 / -	0.15 / -	11.83 / -	27.69 / -	5.33 / -	0.85 / -
	-	VFMM3D(Voxel-RCNN)	-	7.48 / 7.24	17.41 / 16.91	2.08 / 1.99	0.28 / 0.24	18.26 / 17.46	38.26 / 36.75	8.35 / 7.95	1.60 / 1.36
	-	VFMM3D(PV-RCNN)	-	8.06 / 7.79	19.38 / 18.81	2.22 / 2.05	0.15 / 0.12	18.39 / 17.48	39.26 / 37.66	8.81 / 8.06	1.10 / 0.87
L2	LiDAR	CaDDN	CVPR 21	4.49 / 4.45	14.50 / 14.38	1.42 / 1.41	0.09 / 0.09	16.51 / 16.28	44.87 / 44.33	8.99 / 8.86	0.58 / 0.55
	LiDAR	MonoNeRD [53]	ICCV 23	10.03 / 9.93	27.75 / 27.48	5.25 / 5.21	0.60 / 0.59	29.29 / 28.84	60.91 / 60.08	25.36 / 25.00	5.77 / 5.66
	LiDAR	DID-M3D [54]	ECCV 22	- / -	- / -	- / -	- / -	19.37 / 19.19	40.77 / 40.46	15.18 / 15.04	4.69 / 4.59
	Depth	PatchNet [55]	ECCV 20	0.38 / 0.36	1.67 / 1.63	0.13 / 0.11	0.03 / 0.03	2.42 / 2.28	10.01 / 9.73	1.07 / 0.94	0.22 / 0.16
	Depth	PCT [56]	NeurIPS 21	0.66 / 0.66	3.18 / 3.15	0.27 / 0.26	0.07 / 0.07	4.03 / 3.99	14.67 / 14.51	1.74 / 1.71	0.36 / 0.35
	-	M3D-RPN [12]	ICCV 19	0.33 / 0.33	1.12 / 1.10	0.18 / 0.17	0.02 / 0.02	3.61 / 3.46	11.12 / 10.67	2.12 / 2.04	0.24 / 0.20
	-	GUPNet [9]	ICCV 21	2.14 / 2.12	6.13 / 6.08	0.78 / 0.77	0.02 / 0.02	9.39 / 9.31	24.69 / 24.50	4.67 / 4.62	0.19 / 0.19
	-	DEVIANT [57]	ECCV 22	2.52 / 2.50	6.93 / 6.87	0.95 / 0.94	0.02 / 0.02	10.29 / 10.20	26.75 / 26.54	4.95 / 4.90	0.16 / 0.16
	-	MonoJSG [58]	CVPR 22	0.91 / 0.89	4.64 / 4.65	0.55 / 0.53	0.09 / 0.09	5.34 / 5.17	20.79 / 20.19	3.79 / 3.67	0.85 / 0.82
	-	SSD-MonoDETR [59]	TIV 23	4.12 / -	8.87 / -	1.02 / -	0.13 / -	11.34 / -	27.62 / -	5.21 / -	0.76 / -
	-	VFMM3D(Voxel-RCNN)	-	6.62 / 6.41	17.17 / 16.67	1.89 / 1.81	0.22 / 0.19	16.19 / 15.48	37.77 / 36.29	7.58 / 7.21	1.25 / 1.06
	-	VFMM3D(PV-RCNN)	-	7.14 / 6.89	19.11 / 18.55	2.01 / 1.86	0.12 / 0.09	16.31 / 15.50	38.78 / 37.19	8.01 / 7.33	0.86 / 0.68

TABLE II

KITTI valにおける我々のモデルと最新モデルの比較。SET FOR CAR CLASS.'m od.'は中程度の難易度レベルを示す。* モデルの深度推定器のトレーニングセットが検証セットと重なる可能性があることを意味する。R ed は最良の結果を示し、Blue は 2 番目の最良の結果を示す。

Method	Reference	3D AP@0.7			BEV AP@0.7			3D AP@0.5			BEV AP@0.5		
		Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
Monopair [6]	CVPR 2020	16.28	12.30	10.42	24.12	18.17	15.76	55.38	42.39	37.99	61.06	47.63	41.92
MonoDLE [16]	CVPR 2021	17.45	13.66	11.68	24.97	19.33	17.01	55.41	43.42	37.81	60.73	46.87	41.89
MonoFlex [24]	CVPR 2021	24.22	17.34	15.13	31.65	23.29	20.02	60.70	45.65	39.91	66.26	49.30	44.42
GUPNet [9]	ICCV 2021	22.76	16.46	13.72	31.07	22.94	19.75	57.62	42.33	37.59	61.78	47.06	40.88
DDMP-3D* [60]	CVPR 2021	28.12	20.39	16.34	-	-	-	-	-	-	-	-	-
CaDNN [17]	CVPR 2021	23.57	16.31	13.84	-	-	-	-	-	-	-	-	-
MonoRUn [26]	CVPR 2021	20.02	14.65	12.61	-	-	-	59.71	43.39	38.44	-	-	-
HomoLoss [61]	CVPR 2022	23.04	16.89	14.90	31.04	22.99	19.84	-	-	-	-	-	-
MonoDDE [62]	CVPR 2022	26.66	19.75	16.72	35.51	26.48	23.07	-	-	-	-	-	-
MonoDTR [27]	CVPR 2022	24.52	18.57	15.51	33.33	25.35	21.68	64.03	47.32	42.20	69.04	52.47	45.90
MonoGround [18]	CVPR 2022	25.24	18.69	15.58	32.68	24.79	20.56	62.60	47.85	41.97	67.36	51.83	45.65
OPA-3D [63]	IRAL 2023	24.97	19.40	16.59	33.80	25.51	22.13	-	-	-	-	-	-
MonoNeRD [53]	ICCV 2023	20.64	15.44	13.99	29.03	22.03	19.41	-	-	-	-	-	-
PDR [4]	TCSVT 2023	27.65	19.44	16.24	35.59	25.72	21.35	-	-	-	-	-	-
VFMM3D(PointPillars)	-	22.43	15.45	13.92	35.85	24.88	22.75	63.52	45.60	41.79	68.26	48.62	44.75
VFMM3D(Voxel-RCNN)	-	29.09	19.41	17.09	39.43	26.56	23.69	68.72	50.53	45.02	73.17	52.90	47.25
VFMM3D(PV-RCNN)	-	29.05	19.10	16.86	41.78	28.53	25.61	69.95	51.78	47.01	74.24	54.98	49.69
VFMM3D(PointPillars)*	-	27.23	17.15	13.95	38.02	24.93	20.79	69.94	48.75	42.98	74.12	54.37	48.09
VFMM3D(Voxel-RCNN)*	-	34.60	21.58	18.23	44.18	28.66	24.02	72.85	51.83	45.53	76.69	55.43	48.95
VFMM3D(PV-RCNN)*	-	32.06	21.00	17.49	43.30	28.88	24.72	73.21	53.51	46.88	77.09	58.37	51.69

このフレームワークは、新しいボクセルセット抽象化モジュールを介して、3Dシーンをキーポイントの小さなセットに要約し、次に、提案洗練のために、これらのキーポイントの特徴をRoI-gridポイントに集約する。このように点ベースとボクセルベースの特徴学習を統合することで、管理可能なメモリ消費で3D物体検出の性能向上を実現する。

PointPillars PointPillarsは、点群から効率的に物体を検出するために設計された手法であり、特に自律走行アプリケーションに有効である。垂直の列や柱に編成された点群の表現を学習するためにPointNetsを使用する新しいエンコーダを紹介する。符号化された特徴は、任意の標準的な2次元畳み込み検出アーキテクチャと統合することができ、PointPillarsはこの統合のために、合理化された下流ネットワークを導入する。

GPU上で垂直方向のビニングを調整し、2次元畳み込みの効率を活用する。本手法は、速度と精度の両方において、従来のエンコーダを大幅に上回り、62Hzで動作し、3Dと鳥瞰図検出の両方において、KITTIベンチマークでの性能に新しい基準を設定した。

IV. 実験

A. Settings

データセット WaymoとKITTIの2つのデータセットで本手法を評価する。Waymoデータセット[64]は、多様な運転環境から収集された1,150のビデオシーケンスを含む。公式分割プロトコルを用いて、データセットを798配列、158081サンプルの訓練セットと202配列、39,848サンプルの検証セットに分割する。

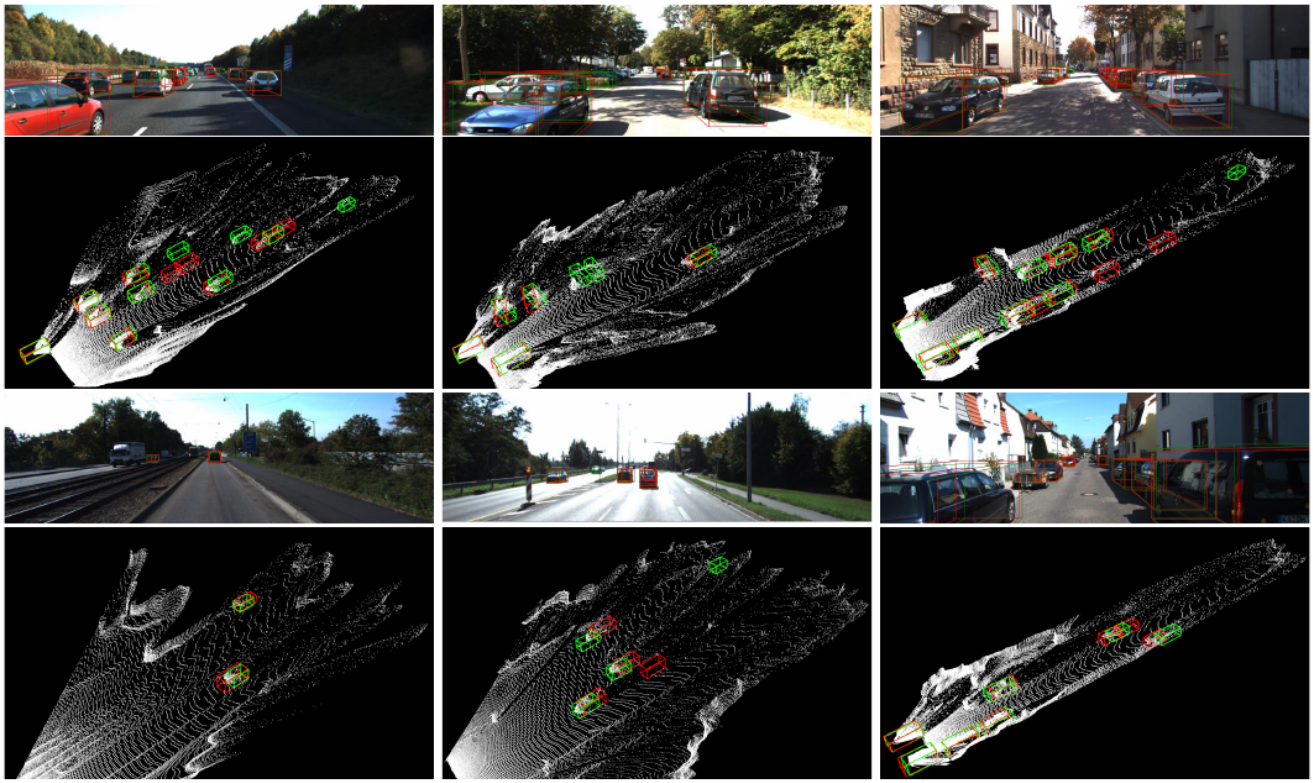

図3. KITTI値に対するVFMM3Dの定性的結果。セット。フロントビュー画像(1行目と3行目)と擬似LiDAR点群(2行目と4行目)において、グランドトゥルースの注釈(緑)と共に3Dバウンディングボックス推定値(赤)を可視化する。

DEVIANT[57]、PCT[56]、および他の手法[53]、[58]に従い、単眼3D物体検出のためにフロントビューカメラを使用する。KITTIデータセットには、3Dオブジェクト検出と鳥瞰図[65]の2つのベンチマークが含まれており、7,481のトレーニングサンプルと7,518のテストサンプル、および対応するLiDAR点群、ステレオ画像、フルカメラマトリックスで構成されている。元の学習サンプルを、従来の手法[29]，[31]に従い、3,712サンプルの訓練用スピルトセットと3,769サンプルの検証用スプリットセットに分割した。トレーニングフェーズとテストフェーズの両方において、我々のアプローチはシングルビューのRGB画像データのみを利用し、LiDAR点群やステレオ画像データは組み込まないことは注目に値する。

評価指標 Waymoでは、公式の評価指標に従い、2つのオブジェクトレベルで評価する：レベル1とレベル2にmAPとmAPHを付与し、各オブジェクトをその3Dバウンディングボックス内に含まれるLiDAR点の数に応じてレベルに割り当てる。さらに、3つの距離範囲での性能も示す：[0m，30m)，[30m，50m)，[50m，∞]．KITTIについては、3つの異なる難易度(簡単、中程度、難しい)の下で、標準的な評価指標AP_{40} [66]を用いて、精度-想起曲線の40の想起位置でサンプリングした平均精度の性能を報告する。幅広い検証のために、3DボックスとBEVボックスの車クラスのIoU閾値を0.7(AP @0.7)と0.5(AP @0.5)の2つ選択した。

トレーニングの詳細。画像サイズはKITTIデータセットでは1224×370ピクセル、Waymoデータセットでは1920×1280ピクセルに設定され、2Dデータ補強は行われない。LiDARベースの検出ステージでは、3Dランダム水平フリップ補強とポイントシャッフル補強を使用する。奥行き推定の精度が低いため、3Dアノテーションボックス内の擬似点と実点の間には大きな乖離がある。したがって、この方法では、GTサンプリング[41]のデータ増強法を利用しない。KITTIデータセットでの実験では、LiDARベースの3D物体検出器としてVoxel-RCNN[43]を使用する。Voxel-RCNNとPVRCNN[44]については、擬似LiDARの範囲をX軸、Y軸、Z軸に沿ってそれぞれ[0，70.4]，[-40.0，40.0]，[-3.0，1.0]mとした。そして、ボクセルサイズは(0.05，0.05，0.1)メートルである。PointPillars[42]では、擬似LiDARの範囲を[0，69.12]，[-39.68，39.68]，[-3.0，1.0]m、柱のサイズを(0.16，0.16，4)mとした。Waymoデータセットでは、Voxel-RCNNとPV-RCNNを使用し、擬似LiDAR範囲を[0，75.2]，[-75.2，75.2]，[-2.0，4.0]メートル、ボクセルサイズ(0.1，0.1，0.15)メートル、[0，59.6]，[-25.6，25.6]，[-2.0，4.0]メートル、ボクセルサイズ(0.05，0.05，0.15)メートルとする。

我々の手法は主にOpenPCDet [67]を用いて実装されている。我々は、Grounding DINO、SAM with vit-h、DAM with vit-lを含む事前に訓練されたモデルを利用する。これは、KITTI屋外メトリック深度データセット上で微調整され、メトリック深度推定能力を含む。しかし、公式に提供されているDAMモデルは、DD3D [30]やPesuo-Lidar++ [50]で述べられているように、KITTI-depthデータセット全体で微調整されており、このトレーニングセットは検出のためのKITTI-3D検証データと重複していることに注意する。そこで、KITTI-3Dのクリーントレーニングスプリットを使用する。

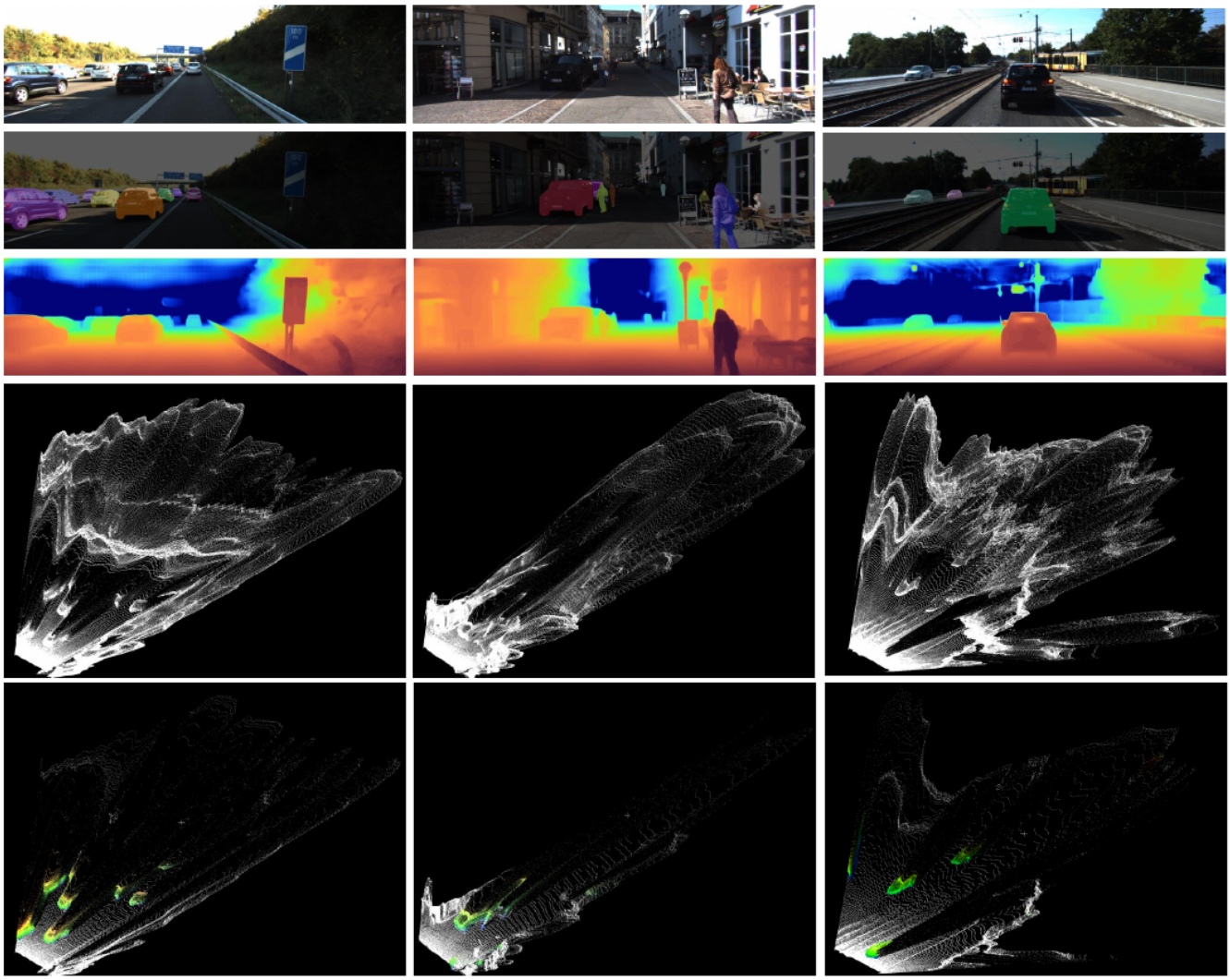

図4. テキストプロンプトSAMによるセグメンテーション結果(2行目)、DAMによる深度マップ(3行目)、深度マップによる生の擬似LiDAR(4行目)、KITTI上のVFMM3Dの各コンポーネントによるPainted Sparse Pseudo-LiDAR(5行目)の結果の可視化。

これは、DD3Dによって提供されたKITTI-3D画像のいずれにも地理的に近いトレーニング画像を削除し、偏りを避けるためにDAMを微調整する。また、KITTI-depthデータセット全体で微調整したDAMを用いたKITTIの結果も示す。いくつかの手法[60]は、指標*を用いたTbale IIで示すように、この種の結果も報告しているからである。LiDARベースの3Dオブジェクト検出器は、8つのRTX A6000 GPUを使用して、スパースペイント擬似LiDAR上で学習される。AdamW[68]オプティマイザを採用し、学習率は4e-4、重み減衰は0.01である。バッチサイズはPointPillarsが32、Voxel-RCNNが32、PV-RCNNが8である。さらに、学習エポック数は全ての検出器で80に設定されている。

B. Waymoの結果

表 IはWaymo検証用スピルトでの性能比較結果である。これは、我々の手法VFMM3Dが、学習中に余分なモダリティデータを必要としない他の単眼手法[9]、[12]、[57]、[58]を凌駕し、最新の単眼手法DEVIANT[57]をレベル1とレベル2の3D mAP(IoU=0.5)でそれぞれ7.41%と6.20%上回ることを示している。注目すべきは、我々の手法が、LiDARを追加学習データとして使用するCaDNN [17]をも上回り、レベル1の3D mAP(IoU=0.5)で0.72%であったことである。

C. KITTI Results

KITTI値に対する本手法の性能を比較する。を用いた車種分類の性能評価を行った.関連する結果をTab. II.本手法は、すべての難易度において、3DとBEVの両方で既存手法を凌駕していることがわかる。具体的には、IoUを0.5に設定した場合、我々の方法はこれらの方法をさらに大きなマージンで上回る。最新の単眼物体検出手法MonoNeRD[53]と比較すると、我々のアプローチは3Dで8.41%、BEVで12.75%、イージーレベルで上回る。KITTI値に対する本手法による3次元検出ボックスの定性的な結果。のセットを可視化したのが図3である。

TABLE III
ABLATION ANALYSIS ON KITTI *val.* SET. WE QUANTIFY THE IMPACTS OF PAINTING STRATEGY, *RGB P* MEANS PAINTING ALL PSEUDO-LIDAR WITH CORRESPONDING RGB PIXEL, *SAM P* REFERS TO PSEUDO-LIDAR PAINTING BY TEXT-PROMPT SAM.

DAM	RGB-P	SAM-P	3D AP@0.7			BEV AP@0.7		
			Easy	Mod.	Hard	Easy	Mod.	Hard
✓			21.50	15.88	14.05	34.81	25.38	22.79
✓	✓		26.96	17.77	15.77	36.63	25.60	23.51
✓		✓	29.05	19.10	16.86	41.78	28.53	25.61

D. 各コンポーネントの結果の可視化

VFMM3Dの各コンポーネントの結果を図4に可視化する。2行目と3行目では、Text-prompt SAMとDAMの両方が、良好なセグメンテーション結果と正確な深度マップを生成している。4行目と5行目を比較すると、我々の設計したスパース化とペインティング操作によって得られたスパース擬似LiDARは、深度マップから直接得られる密な擬似LiDARとは対照的に、前景点の情報を高めるときに、無駄なノイズ点を大幅に減らすことができることが明らかである。

E. アブレーション研究

本節では、擬似LiDAR絵画戦略の影響と、異なるLiDARベースの3D物体検出器のアブレーション研究という2つの重要な観点から、我々の手法を掘り下げる。

塗装戦略の違いによる影響PV-RCNN3D検出器を用いて、異なる擬似LiDAR絵画戦略が最終的な3D検出結果に与える影響を評価するために、アブレーション実験を行う。表IIIに示すように、我々は、テキストプロンプトSAMによって提供される詳細で正確な前景情報(SAM-P)が、後続のLiDARベースの3D検出器を効率的に集中させることを観察する。グローバルペインティングアプローチ(RGB-P)と比較すると、3DとBEVの検出結果に対して、それぞれ1.51%と3.39%全体の平均検出精度が向上している。異なるLiDARベースの3Dオブジェクトに対するアブレーション研究
検出器。VFMM3Dの汎用性を示すために、PointPillars [42]、PV-RCNN [44]、VoxelRCNN [43]など、様々なLiDARベースの3D物体検出器を採用する。異なる3D検出器が予測精度に与える影響をTab. IIとTab. I.

V. CONCLUSION

本論文では、Vision Foundation Models (VFM)を活用し、3D物体検出のための単眼画像データの可能性を解き放つ、単眼3D物体検出のための新しいフレームワークであるVFMM3Dを紹介した。VFMM3Dは、セグメント何でもモデル(SAM)とデプス何でもモデル(DAM)を統合することで、特定のデータセットで微調整を行うことなく、詳細な空間情報で深度特徴を豊かにする擬似LiDARデータを生成する。我々の提案するアプローチは、計算効率やノイズの低減など、単眼3D物体検出における主要な課題に対処する。VFMM3Dは、多様なLiDARベースの検出器にシームレスに統合することができ、検出アーキテクチャ全体への汎用性と適用性を高めることができる。

WaymoデータセットとKITTIデータセットを用いた広範な実験を通して、様々な難易度において、VFMM3Dが既存の手法よりも優れた性能を持つことを実証した。我々のフレームワークは、一貫して最先端のアプローチを凌駕し、単眼画像から正確な3Dオブジェクト表現を抽出する有効性を示す。この結果は、自律走行やロボット工学のアプリケーションにおける実世界展開のための、ロバストで適応性のあるソリューションとしてのVFMM3Dの可能性を強調するものである。

今後の研究では、屋内や複雑な気象条件など、様々な実世界のシナリオにおける追加データセットを調査することで、VFMM3Dの能力をさらに探求する。さらに、検出精度を損なうことなく計算効率を向上させるために、スパース化技術を改良する予定である。

REFERENCES

[1] L. Zhao, J. Guo, D. Xu, and L. Sheng, "Transformer3d-det: Improving 3d object detection by vote refinement," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 12, pp. 4735–4746, 2021.

[2] M. Song, S. Lim, and W. Kim, "Monocular depth estimation using laplacian pyramid-based depth residuals," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 11, pp. 4381–4393, 2021.

[3] S. Chen, Z. Pu, X. Fan, and B. Zou, "Fixing defect of photometric loss for self-supervised monocular depth estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1328–1338, 2022.

[4] H. Sheng, S. Cai, N. Zhao, B. Deng, M.-J. Zhao, and G. H. Lee, "Pdr: Progressive depth regularization for monocular 3d object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 12, pp. 7591–7603, 2023.

[5] Y. Cao, Z. Wu, and C. Shen, "Estimating depth from monocular images as classification using deep fully convolutional residual networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 11, pp. 3174–3182, 2018.

[6] Y. Chen, L. Tai, K. Sun, and M. Li, "Monopair: Monocular 3d object detection using pairwise spatial relationships," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[7] J. Ku, A. D. Pon, and S. L. Waslander, "Monocular 3d object detection leveraging accurate proposals and shape reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[8] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka, "3d bounding box estimation using deep learning and geometry," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 7074–7082.

[9] Y. Lu, X. Ma, L. Yang, T. Zhang, Y. Liu, Q. Chu, J. Yan, and W. Ouyang, "Geometry uncertainty projection network for monocular 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 3111–3121.

[10] X. Shi, Q. Ye, X. Chen, C. Chen, Z. Chen, and T.-K. Kim, "Geometry-based distance decomposition for monocular 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 172–15 181.

[11] Z. Liu, D. Zhou, F. Lu, J. Fang, and L. Zhang, "Autoshape: Real-time shape-aware monocular 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 641–15 650.

[12] G. Brazil and X. Liu, "M3d-rpn: Monocular 3d region proposal network for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[13] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3d object detection for autonomous driving," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2147–2156.

[14] A. Simonelli, S. R. Bulo, L. Porzi, M. López-Antequera, and P. Kontschieder, "Disentangling monocular 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1991–1999.

[15] Y. Zhang, W. Zheng, Z. Zhu, G. Huang, D. Du, J. Zhou, and J. Lu, "Dimension embeddings for monocular 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1589–1598.

[16] X. Ma, Y. Zhang, D. Xu, D. Zhou, S. Yi, H. Li, and W. Ouyang, "Delving into localization errors for monocular 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 4721–4730.

[17] C. Reading, A. Harakeh, J. Chae, and S. L. Waslander, "Categorical depth distribution network for monocular 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 8555–8564.

[18] Z. Qin and X. Li, "Monoground: Detecting monocular 3d objects from the ground," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 3793–3802.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[20] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2024.

[21] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[22] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[23] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 9650–9660.

[24] Y. Zhang, J. Lu, and J. Zhou, "Objects are different: Flexible monocular 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 3289–3298.

[25] T. Wang, Z. Xinge, J. Pang, and D. Lin, "Probabilistic and geometric depth: Detecting objects in perspective," in *Conference on Robot Learning*. PMLR, 2022, pp. 1475–1485.

[26] H. Chen, Y. Huang, W. Tian, Z. Gao, and L. Xiong, "Monorun: Monocular 3d object detection by reconstruction and uncertainty propagation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 10 379–10 388.

[27] K.-C. Huang, T.-H. Wu, H.-T. Su, and W. H. Hsu, "Monodtr: Monocular 3d object detection with depth-aware transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 4012–4021.

[28] R. Zhang, H. Qiu, T. Wang, Z. Guo, Z. Cui, Y. Qiao, H. Li, and P. Gao, "Monodetr: Depth-guided transformer for monocular 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 9155–9166.

[29] X. Ma, Z. Wang, H. Li, P. Zhang, W. Ouyang, and X. Fan, "Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6851–6860.

[30] D. Park, R. Ambrus, V. Guizilini, J. Li, and A. Gaidon, "Is pseudo-lidar needed for monocular 3d object detection?" in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 3142–3152.

[31] X. Weng and K. Kitani, "Monocular 3d object detection with pseudo-lidar point cloud," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[32] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8445–8453.

[33] L. Liu, J. Lu, C. Xu, Q. Tian, and J. Zhou, "Deep fitting degree scoring network for monocular 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1057–1066.

[34] P. Li, H. Zhao, P. Liu, and F. Cao, "Rtm3d: Real-time monocular 3d detection from object keypoints for autonomous driving," in *European Conference on Computer Vision*. Springer, 2020, pp. 644–660.

[35] S. Shi, X. Wang, and H. Li, "Pointrcnn: 3d object proposal generation and detection from point cloud," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. IEEE, 2019, pp. 770–779.

[36] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

[37] Z. Yang, Y. Sun, S. Liu, X. Shen, and J. Jia, "Std: Sparse-to-dense 3d object detector for point cloud," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1951–1960.

[38] D. Zhou, J. Fang, X. Song, L. Liu, J. Yin, Y. Dai, H. Li, and R. Yang, "Joint 3d instance segmentation and object detection for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1839–1849.

[39] Q. Wang, J. Chen, J. Deng, and X. Zhang, "3d-centernet: 3d object detection network for point clouds with center estimation priority," *Pattern Recognition*, vol. 115, p. 107884, 2021.

[40] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4490–4499.

[41] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.

[42] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. IEEE, 2019, pp. 12 697–12 705.

[43] J. Deng, S. Shi, P. Li, W. Zhou, Y. Zhang, and H. Li, "Voxel r-cnn: Towards high performance voxel-based 3d object detection," in *Proceedings of the AAAI conference on artificial intelligence*, 2021, pp. 1201–1209.

[44] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "Pv-rcnn: Point-voxel feature set abstraction for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. IEEE, 2020, pp. 10 529–10 538.

[45] X. Zhao, W. Ding, Y. An, Y. Du, T. Yu, M. Li, M. Tang, and J. Wang, "Fast segment anything," *arXiv preprint arXiv:2306.12156*, 2023.

[46] L. Ke, M. Ye, M. Danelljan, Y.-W. Tai, C.-K. Tang, F. Yu *et al.*, "Segment anything in high quality," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[47] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "Pointpainting: Sequential fusion for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4604–4612.

[48] S. Xu, D. Zhou, J. Fang, J. Yin, Z. Bin, and L. Zhang, "Fusionpainting: Multimodal fusion with adaptive attention for 3d object detection," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021, pp. 3047–3054.

[49] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.

[50] Y. You, Y. Wang, W.-L. Chao, D. Garg, G. Pleiss, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=BJedHRVtPB

[51] H. Zhu, J. Deng, Y. Zhang, J. Ji, Q. Mao, H. Li, and Y. Zhang, "Vpfnet: Improving 3d object detection with virtual point based lidar and stereo data fusion," *IEEE Transactions on Multimedia*, 2022.

[52] H. Wu, C. Wen, S. Shi, X. Li, and C. Wang, "Virtual sparse convolution for multimodal 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 653–21 662.

[53] J. Xu, L. Peng, H. Cheng, H. Li, W. Qian, K. Li, W. Wang, and D. Cai, "Mononerd: Nerf-like representations for monocular 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 6814–6824.

[54] L. Peng, X. Wu, Z. Yang, H. Liu, and D. Cai, "Did-m3d: Decoupling instance depth for monocular 3d object detection," in *European Conference on Computer Vision*. Springer, 2022, pp. 71–88.

[55] X. Ma, S. Liu, Z. Xia, H. Zhang, X. Zeng, and W. Ouyang, "Rethinking pseudo-lidar representation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*. Springer, 2020, pp. 311–327.

[56] L. Wang, L. Zhang, Y. Zhu, Z. Zhang, T. He, M. Li, and X. Xue, "Progressive coordinate transforms for monocular 3d object detection,"

Advances in Neural Information Processing Systems, vol. 34, pp. 13 364–13 377, 2021.

[57] A. Kumar, G. Brazil, E. Corona, A. Parchami, and X. Liu, "Deviant: Depth equivariant network for monocular 3d object detection," in *European Conference on Computer Vision*. Springer, 2022, pp. 664–683.

[58] Q. Lian, P. Li, and X. Chen, "Monojsg: Joint semantic and geometric cost volume for monocular 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 1070–1079.

[59] X. He, F. Yang, K. Yang, J. Lin, H. Fu, M. Wang, J. Yuan, and Z. Li, "Ssd-monodetr: Supervised scale-aware deformable transformer for monocular 3d object detection," *IEEE Transactions on Intelligent Vehicles*, 2023.

[60] L. Wang, L. Du, X. Ye, Y. Fu, G. Guo, X. Xue, J. Feng, and L. Zhang, "Depth-conditioned dynamic message propagation for monocular 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 454–463.

[61] J. Gu, B. Wu, L. Fan, J. Huang, S. Cao, Z. Xiang, and X.-S. Hua, "Homography loss for monocular 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 1080–1089.

[62] Z. Li, Z. Qu, Y. Zhou, J. Liu, H. Wang, and L. Jiang, "Diversity matters: Fully exploiting depth clues for reliable monocular 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 2791–2800.

[63] Y. Su, Y. Di, G. Zhai, F. Manhardt, J. Rambach, B. Busam, D. Stricker, and F. Tombari, "Opa-3d: Occlusion-aware pixel-wise aggregation for monocular 3d object detection," *IEEE Robotics and Automation Letters*, vol. 8, no. 3, pp. 1327–1334, 2023.

[64] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2446–2454.

[65] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[66] A. Simonelli, S. R. Bulo, L. Porzi, M. Lopez-Antequera, and P. Kontschieder, "Disentangling monocular 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[67] O. D. Team, "Openpcdet: An open-source toolbox for 3d object detection from point clouds," https://github.com/open-mmlab/OpenPCDet, 2020.

[68] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.